DOCUMENT RESUME

ED 318 743 TM 014 650

AUTHOR Jaeger, Richard M.
TITLE Establishing Standards on Tests Used for Certification of Education Personnel: Validity Issues.
PUB DATE Aug 88
NOTE 23p.; Paper presented at the Annual Meeting of the American Psychological Association (96th, Atlanta, GA, August 12-16, 1988) and, with slight modifications, under the title "Setting and Using Cutoff Scores on Tests Used for Certification" at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; Elementary Secondary Education; Evaluators; *Licensing Examinations (Professions); *School Personnel; Scoring; *Standards; *Teacher Certification; Teacher Selection; Test Use; *Validity
IDENTIFIERS *Standard Setting; Standards for Educational and Psychological Tests; Uniform Guidelines Employee Selection Procedures

ABSTRACT

The "Standards for Educational and Psychological Testing" (1985) and the "Uniform Guidelines on Employee Selection Procedures" (1978) provide no clear guidance on choosing methods for setting standards on teacher certification tests, although they require that practices be well-documented and produce reasonable results, and that the implications of conclusions be validated. Although literature to guide choices is lean, it appears that Angoff's procedure produces more reasonable standards than do its competitors. Practices that are recommended for use include: (1) identification of a population of judges qualified to make judgments; (2) use of sound scientific methods to sample judges; (3) selection of samples of sufficient size to provide stable average recommendations; (4) use of standard-setting procedures within the capabilities of judges; (5) training judges to understand and interpret information; (6) informing judges about requirements of the teaching jobs in question; (7) providing judges with information on the relationship between test and job performance; (8) providing judges with information on the distribution of test performances of a representative sample of examinees; and (9) providing judges with opportunities to reconsider initial recommendations. Following these guidelines will yield test standards produced by an appropriate and adequate sample of capable and qualified judges. A 25-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
RICHARD M. JAEGER
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED318743

Establishing Standards on Tests Used for Certification of Education Personnel: Validity Issues

Richard M. Jaeger
University of North Carolina at Greensboro

Presented at the 96th Annual Convention of the
American Psychological Association at
Atlanta, Georgia
August 1988

(A Modification of a paper entitled
"Setting and Using Cutoff Scores on Tests Used for Certification"
presented at the Annual Meeting of the
National Council on Measurement in Education
New Orleans, April, 1988)

TM014650

BEST COPY AVAILABLE

The Purpose and Meaning of Teacher Certification Tests

According to the 1985 *Standards for Educational and Psychological Testing*:

> The primary purpose of licensure or certification is to protect the public. Licensing requirements are imposed to ensure that those licensed possess knowledge and skills in sufficient degree to perform important occupational activities safely and effectively. (p. 63).

Although state education agencies have traditionally issued "teaching certificates," such certificates have played the role of licenses, since without a valid certificate, one cannot legally teach in the public schools of the issuing state. By similar logic, what various state education agencies have labeled "certification tests" for teachers and other certificated personnel actually constitute licensure tests, since typically, an examinee must earn a passing score as a prerequisite to certification.

In keeping with the *1985 Standards* definition of licensure tests, most teacher certification tests are designed to assess some aspect of examinees' minimum qualifications to teach. They are not designed to measure teaching proficiency or typical teaching performance. In fact, most teacher certification tests assess examinees' knowledge in a limited range of subject matter; sometimes including knowledge of pedagogical technique but more often focused on subject-matter knowledge in a specific area of certification, such as elementary education, social studies, language arts, mathematics, etc.

Competence is a Continuum but Certification is Binary

Even when restricted to specific subject-matter knowledge, it is generally recognized that the construct to be measured by a teacher certification test is a continuous variable Glass (1978), Jaeger (1986), Linn, Madaus & Pedulla (1982), Shepard (1979, 1980a, 1980b, 1984). There is no Platonically-determined point on the continuum, above which an examinee is competent to teach, and below which

an examinee is incompetent. Nonetheless, certification decisions *are* binary; the state, as a licensing agency, either grants or withholds a certificate to teach, depending in part on a candidate's performance on a teacher certification test. It is the necessity of making a certification decision that impels the state to translate an assessment of status on a continuous construct into a binary classification.

Past critics of competency testing (cf., Glass, 1978) have condemned the practice of setting performance standards on the basis of their arbitrariness and the false assumption that users of competency tests implicitly deny the existence of an underlying continuum of competence. The latter charge is unsupported by fundamental logic or evidence. The former certainly applies to certification test standards that have been established by some state agencies, but is not implicit in the process of setting standards. If carefully applied, methods for establishing certification test standards can themselves satisfy accepted standards of the measurement profession.

Requirements of the *1985 Test Standards* and the *Uniform Guidelines*

The 1985 *Test Standards* impose explicit requirements, and additional requirements through their general provisions, on those who would establish standards on teacher certification tests. That these *Standards* apply to agencies that develop and use teacher certification tests is indisputable. The introduction to the *Test Standards* (p. 3) defines tests as including "... standardized ability (aptitude and achievement) instruments, diagnostic and evaluative devices, interest inventories, personality inventories, and projective instruments. And

Chapter 11 of the *Test Standards* is entitled "Professional and Occupational Licensure and Certification."

The first sentence of Standard 3.1 (p. 25) imposes the most pervasive and perhaps the most controversial requirement that applies to the establishment of standards on teacher certification tests. It states that "Tests and testing programs should be developed on a *sound scientific basis*." (emphasis added). Although broadly interpretable, the fundamental canons of science would require that procedures for establishing test standards be well documented, be based in an explicable rationale, be public, be replicable, and be capable of producing a reliable result.

More explicitly, Standard 10.9 requires that any technical basis for the establishment of test standards be clearly explained; Standard 6.9 requires that the method and rationale that underlie any test standard be reported, and when test standards are based on professional judgment, that qualifications of the judges be reported.

The applicability of the 1978 *Uniform Guidelines on Employee Selection Procedures* to teacher certification tests developed and used by state governments is subject to legal dispute. Rebell (1986) noted that the relevant case law is inconsistent. The major issue appears to turn on whether the state would be considered an "employer" within the meaning of Title VII of the Civil Rights Act of 1964. In light of the inconsistent position of the courts, it would seem prudent to assume that the *Uniform Guidelines* are applicable to teacher certification testing programs, and to incorporate compliance with the requirements of the *Guidelines* into the design of such programs.

The *Uniform Guidelines* are more explicit in their requirements for setting standards on employee selection tests than are the *1985 Test Standards*. The *Uniform Guidelines* mirror the documentation requirements of the *Test Standards*, but require in addition, examination of the psychometric properties of the standard-setting procedure employed and consideration of the consequences of using the resulting standards. The *Uniform Guidelines* specifically require the following:

> Users should evaluate each selection procedure to assure that it is appropriate for operational use, including establishment of cutoff scores or rank ordering (p. 38301).
>
> The methods considered for use of the selection procedure (e.g., as a screening device with a cutoff score ...) and available evidence of their impact should be described (essential). This description should include the rationale for choosing the method for operational use, and evidence on the validity and utility of the procedure as it is to be used (essential) (p. 38306).
>
> If the selection procedure is used with a cutoff score, the user should describe the way in which normal expectations of proficiency within the work force were determined and the way in which the cutoff score was determined (p. 38306).
>
> Where cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force. Where applicants are ranked on the basis of properly validated selection procedures and those applicants scoring below a higher cutoff score than appropriate in light of such expectations have little or no chance of being selected for employment, the higher cutoff score may be appropriate, but the degree of adverse impact should be considered (p. 38298).

In summary, the specific provisions of the *1985 Test Standards* and the *Uniform Guidelines* impose few operational constraints on methods used to establish standards for teacher certification tests. Both require that the method

used be well documented and that a clear rationale for selecting the method be provided. The *Uniform Guidelines* impose the additional requirements that the resulting standards be "reasonable," and that the effects of using the standards, in terms of their impact, utility, and potential adverse impact, be examined and reported. Finally, the *Uniform Guidelines* require that the validity of the standard-setting method be reported.

The "sound scientific basis" criterion of the *1985 Test Standards* can be interpreted to impose the additional requirement that the resulting standards be reliable.

Implications of the *1985 Test Standards* and the *Uniform Guidelines*

Selection of a standard-setting method would be greatly simplified if the requirements of the *Test Standards* or the *Uniform Guidelines* led to the clear exclusion of some methods or to clear preference for others. But such implications are not obvious. The documentation requirements of the *Test Standards* and the *Uniform Guidelines* can be met readily for almost any standard-setting method, whether reasonable or not. And the requirement that a rationale be described makes greater demands on the glibness of the proposer than on the reasonableness of a standard-setting method. We are left, then, to consider the reasonableness and validity reporting criteria imposed by the *Uniform Guidelines* and the reliability requirement deduced from the "sound scientific basis" criterion of the *Test Standards*.

Reasonableness. Logic dictates that standards established for licensure tests should be consistent with the fundamental purpose of such tests -- to protect the public from unqualified would-be practitioners. In the case of teacher

certification tests, the public should be protected from those who are grossly unqualified to teach by virtue of their lack of knowledge of the subject-matter assessed by the test, provided the job relevance of the test content has been clearly established. The reasonableness criterion of the *Uniform Guidelines* derives from this logic. How should it be tested? If standards are to be established for tests used in screening applicants for initial teacher certification, it should be the case that virtually all qualified practicing teachers should exhibit test performance that exceeds the standard. Of course, the rub is the word "qualified." If the test standard were established at, say, the first percentile of the distribution of test scores of a representative sample of practicing teachers, the procedure would imply that virtually all practicing teachers were qualified. Both the public and the legislatures that require teacher certification tests would reject that assumption and would demand a more stringent standard. It is tempting to suggest that the test performances of virtually all teachers nominated as satisfactory by their peers or their administrators, or that the test performances of virtually all teachers whose students average greater than expected progress in achievement during the preceding academic year, should exceed a "reasonable" test standard. This requirement either places trust in the judgment of teachers' peers and administrators, or shifts the burden of defining acceptable performance from a focus on the test performance of would-be teachers to the achievement performance of enrolled students.

To the best of my knowledge, these procedures for investigating the "reasonableness" of standards produced by various standard-setting methods have not been applied. A number of studies in the standard-setting literature confirm that various methods produce different results when they are applied by randomly

equivalent samples of judges or even the same sample of judges; cf. Cross, Impara, Frary & Jaeger (1984) for an example of such a study, or Jaeger (in press) for a review of the literature on comparative standard-setting studies. We know from this literature that Nedelsky's (1954) procedure often produces lower standards than does Angoff's (1971) procedure; that Ebel's (1972) procedure often produces higher standards than does Nedelsky's procedure or Angoff's procedure; and that Jaeger's procedure appears to produce higher standards than do Angoff's or Nedelsky's procedures. These results derive from the application of standard-setting procedures to basic skills tests for elementary and secondary school students, and might not generalize to certification tests for teachers. Furthermore, the comparative literature on standard setting tells us nothing about the production of "reasonable" standards, at least by the criteria described earlier.

Reliability. The appropriate indicator of the reliability of a standard-setting procedure is the extent to which it produces consistent classification of an examinee as "competent" or "incompetent" when it is applied to different samples of items from the domain of generalization, by different samples of judges, on different occasions of judgment. Judgments on the comparative reliabilities of various standard-setting procedures can be made with confidence only when we have amassed distributions of decision consistency indices for each, as these methods are used with a variety of tests in a variety of settings. It is not clear *a priori* that the same procedures will be most reliable when applied to various types of certification and licensure tests or when applied to competency tests for students. Different types of judges, who will have differing levels of knowledge of the abilities of examinees in the domain of the test for which a standard is desired, are appropriate in these various applications. It is likely that the

consistency of classifications produced by a standard-setting method will be sensitive to the knowledge judges hold. Several studies confirm that the variability of judges' recommendations is influenced by their ethnic background (Jaeger, 1982), their grading practices, their attitudes toward competency testing (Secolsky, 1987), and their own achievement levels (Saunders, Ryan & Huynh, 1981).

The existing literature has defined the reliability of various standard-setting procedures in terms of the generalizability of mean recommended test standards across samples of test items and samples of judges. Variance due to the inconsistency of judges on different judgment occasions has been ignored.

Although it would be most desirable to have information on the consistency of classifications produced by various standard-setting procedures, Kane and Wilson (1984) have shown that the probability of misclassification is a monotonic function of the magnitude of the standard error of measurement of the test for which a standard is to be established and the variability of mean recommended standards across samples of items and judges. Thus standard-setting procedures that produce smaller variances of mean recommended standards can be expected to produce more reliable standards, in the appropriate decision-consistency sense described earlier.

The empirical literature on the generalizability of mean recommended standards for various standard-setting procedures is sparse. Brennan and Lockwood (1980) applied a two-factor random-effects ANOVA model to data resulting from the application of Angoff's (1971) procedure and Nedelsky's (1954) procedure to "126 four-alternative items in a health-related area" (p. 221) by five raters. They found that the estimated variance of the mean recommended test

standard produced by Nedelsky's method was almost twice as large as that produced by Angoff's method. The difference in results observed by Brennan and Lockwood probably underestimates the true difference, because their ANOVA model assumes that actual item difficulty values and judges' estimates of item difficulty values for minimally competent examinees do not covary, despite their findings of positive correlations that were higher for Angoff's method than for Nedelsky's method. Kane and Wilson (1984) show that the variance of the mean recommended test standard is inversely related to the magnitude of a positive covariance.

Cross, et al. (1984) compared the generalizability of standards produced by modifications of the Angoff and Nedelsky standard-setting procedures and by the Jaeger (1982) standard-setting procedure, when applied to the area examinations in elementary education and mathematics of the National Teacher Examinations. The ANOVA model they used was identical to that employed by Brennan and Lockwood. The results obtained by Cross, et al. varied, depending on the metric used to express reliability. An index of consistency of ratings due to Ebel (1951) was identical for the Angoff and Jaeger procedures, and was lower for Nedelsky's method, as these methods were originally proposed, when applied to both examinations. However, modification of the Angoff and Nedelsky procedures by incorporating iteration and use of normative data on examinees' test performances resulted in a higher consistency index for Angoff's method. The consistency of findings by Brennan and Lockwood (1980) and by Cross, et al. (1984) would be comforting, had the statistic used by the former authors produced the same result when used by the latter authors. However, such was not the case. When applied as they were originally proposed, Cross, et al. found that Angoff's method resulted in

a substantially smaller variance of the mean recommended test standard than did Nedelsky's method for the mathematics test, but a larger variance for the elementary education test. Following iteration and feedback of normative information, the Angoff and Nedelsky methods produced equal variances of the mean recommended standard for the mathematics test, and Nedelsky's method resulted in a variance that was less than half as large for the elementary education test. The variance produced by Jaeger's method was twice as large as that resulting from Angoff's and Nedelsky's methods for the mathematics test, but was slightly smaller than that produced by Angoff's method for the elementary education test.

In a comparison of Angoff's method and a modification of Nedelsky's method, Reilly, Zink & Israelski (1984) found that Angoff's method resulted in somewhat higher reliability estimates for mean recommended standards than did the modified Nedelsky method. Analyses of variance components attributable to standard-setting procedures, raters, items, and interaction between methods and items revealed that variance due to raters dominated all other sources of systematic variance. The tests used in this study were qualification examinations for building maintenance specialists and auto mechanics.

Van der Linden (1982) defined a third component of error variance in standard setting, which he termed "intrajudge inconsistency" (p. 296). He developed an index that reflects the degree to which individual judges are consistent in their judgments of the likelihood that minimally competent examinees will answer test items correctly. The validity of the index depends on the goodness of fit of the test items to an item response theory model. When applied to nine judges who provided recommended standards for a high school physics test, Van der Linden

found that his intrajudge consistency index was higher for Angoff's method than for Nedelsky's method.

In sum, limited empirical evidence on the comparative reliability of various standard-setting methods suggests that Angoff's method, more often than not, yields standards that are more reliable than those produced by Nedelsky's method. However, the available data base is very small; methods used to assess reliability differ across studies, do not take account of some sources of variance, and probably overestimate the variance of mean recommended standards because they assume that different sources of systematic variance do not covary.

Validity. According to the *Test Standards*:

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of specific inferences made from test scores. (p. 9).

In the case of setting standards on teacher certification tests, the obvious "specific inference" of interest is that candidates who score at or above the standard on the certification test are competent and those who score below the standard are not. It is undoubtedly with this inference in mind that Shepard (1980) made the following comment concerning the validity of standards:

> The final recommendation for dealing with subjective and varied standards is to collect validity evidence. Several of the major categories of method that follow [standard-setting methods that are based on judgments about groups, such as the contrasting-groups method and the borderline group method (Zieky and Livingston, 1977)] can be thought of as alternative strategies for providing different insights to the standards problem *or* as validation strategies to

> reflect on how sensible the cutoff scores are that have been based on a logical study of the test. (p. 454).

This recommendation on validating a test standard suggests that there is some truth to the score-based inference that those whose performance is above the standard are competent and those who are below the standard are incompetent, in the sense of being qualified to teach. Yet Shepard acknowledges that:

> Individuals immediately on either side of the standard will be virtually indistinguishable from one another. With a good test, valid distinctions can be made between those who are well above or well below the standard; but pass-fail distinctions near the cutoff will have poor validity because a continuum of performance has been "arbitrarily" dichotomized. (p. 448).

It is most assuredly the case that the distributions of performance on any imaginable criterion variable for those whose teacher certification test scores were close to any specified test standard would be close to identical. Terming a group with scores just above the standard "competent" and those with scores just below the standard "incompetent" would not be supported by the results of any conventional predictive validity study. If the regression of criterion-variable scores on teacher certification test scores had a non-zero slope, it is also true that the distributions of scores on the criterion variable of groups that had markedly different means on the certification test (presumably on each side of the certification test standard) would be largely non-overlapping. But such predictive evidence would confirm nothing more than a linear relationship between performance on the teacher certification test and performance on some valued criterion variable. The evidence would contribute to a claim that the teacher

certification test was valid, but it would contribute nothing to validation of the claim that those with scores above the test standard should be termed "competent" and those with scores below the test standard should be termed "incompetent."

Unfortunately, the additional types of "validity" evidence recommended by Shepard (1980) are unlikely to be practically useful in validating a standard on a teacher certification test. Application of the contrasting-groups method of standard setting would require that a group of "incompetent" teacher certification applicants be identified by some criterion other than their performance on the certification test. In most states, all such applicants are graduates of state-accredited teacher education programs, and have been labeled by their colleges and universities as competent to begin teaching. To identify such graduates as incompetent, a teacher educator most likely to have evidence on the qualifications of the applicant would have to challenge the conclusion of his/her teacher preparation institution.

Normative information amassed for a relevant criterion group would add to our assurance that a test standard was reasonable (an assurance that should not be lightly dismissed), but could not confirm the validity of the fundamental score-based inference implied by the establishment of a specific test standard. For example, a finding that, say, only 20 percent of practicing teachers earned teacher competency test scores that fell below the test standard might be comforting, but it would not validate the point chosen to divide a continuum of test performance into two regions labeled "competent" and "incompetent."

It is my view that we must alter our fundamental conception of validation, as it applies to the establishment of test standards. No conventional test validation

procedure will provide evidence that any score-based dichotomization of the ability scale into two categories labeled "competent" and "incompetent" is correct. We know that the dichotomization is judgmentally based, arbitrary, and wherever placed on the ability scale, will not result in reliable differences in distributions of performance on any valued criterion for groups adjacent to the point of dichotomization. The best we can hope is that those few conventional validation procedures that appear applicable to standards established for teacher certification tests provide evidence that our arbitrariness has not produced results that are markedly inconsistent with other bases for asserting the competence of those who would be teachers and those who currently practice.

I would suggest that we judge the validity of a standard-setting procedure by conceptualizing the universe score (Cronbach, Gleser, Nanda & Rajaratnum, 1972) that would result if ideal conditions of judgment were enjoyed by an ideal population of appropriate judges. In the case of teacher certification tests, a universe score is the test standard that would be recommended by an infinitely large population of judges if that population were (1) fully informed about the job requirements of beginning teachers, (2) had complete information about the relationship between examinees' performances on the teacher certification test and their later performances on the job, and (3) had perfectly reliable information about the distribution of certification test performances of the population of examinees. If there are several populations of qualified judges, there will be several applicable universe scores.

To assess the validity of a procedure for establishing a standard on a teacher certification test, I would suggest that the elements of the procedure be compared to the conditions that would generate a universe score. More specifically, a

proposed standard-setting procedure should be examined from the following perspectives:

a. Has a population of appropriate judges been clearly identified and justified?

b. Has the population of judges been sampled in a way that will provide panels of judges that are representative of the population and sufficient in size to produce stable average recommendations?

c. Either by virtue of their expertise or through the standard-setting procedure itself, has the sample of judges been well informed on (i) the requirements of the job for which examinees are seeking certification, (ii) evidence on the relationship between performance on the certification test and performance on the job, and (iii) the distribution of certification test performance of a representative sample of applicants?

d. Has the judgment task imposed by the standard-setting procedure been designed so that it is within the capabilities of the sampled judges? In particular, is there evidence that (i) the judges understand the task they are to perform, (ii) the judges have been provided with the information they need to perform that task, and (iii) the judges know how to use that information to perform their task?

e. When used operationally, does the standard-setting procedure produce results that are sufficiently stable across judgment occasions, test forms, and samples of judges?

In the face of our admitted arbitrariness, a standard-setting procedure that satisfies these criteria is, perhaps, the best to which we can aspire. The results it would produce should be reasonable; would, by definition, be reliable; and would generalize to the results we could expect, were we able to elicit a test standard under ideal conditions.

A Reprise and some Deductions

The *1985 Test Standards* and the *Uniform Guidelines* provide no clear guidance on a choice among methods for setting standards on teacher certification tests. The *Test Standards* require that we document our practices well, and that we be "scientific" in our endeavors. The *Uniform Guidelines* require in addition, that our practices produce reasonable results and that we validate the implications of our conclusions.

The empirical literature that would guide our choices among methods for setting standards on teacher certification tests is unfortunately lean. Most empirical studies are based on data collected for other testing applications, including high school competency tests or qualification tests in the health professions. Nonetheless, some tentative conclusions can be drawn. It appears that Nedelsky's procedure often results in standards that are substantially lower than those produced by Angoff's procedure and that Ebel's procedure sometimes produces standards that are unreasonably high. In at least one study, Jaeger's procedure produced standards that were markedly higher than those resulting from the use of Angoff's procedure and Nedelsky's procedure. There appears to be a developing consensus, at least in applications outside the health professions, that Angoff's procedure produces more reasonable standards than do its competitors. Based on very limited results, it appears that Angoff's procedure will often

produce more stable (and hence more reliable) standards, than will its competitors. Unfortunately, there is no trustworthy empirical literature on the validity of alternative standard-setting methods. The correlational evidence on relationships between empirical and judgmentally-proposed item difficulties, advanced in the name of validity evidence by Cross, et al. (1984) and Reilly, Zink & Israelski (1984), among others, is weakly applicable to the question of intrajudge consistency, but provides no information on the validity of test standards.

What practice can be recommended? More-recent empirical findings and additional thought have not changed materially the advice offered by several researchers in the past (cf. Shepard, 1980; Jaeger, 1982; Cross, et al., 1984):

1. Identify at least one population of judges that is qualified by virtue of experience with the requirements of the job and knowledge of the capabilities of examinees, to make recommendations on appropriate test standards.
2. Use sound scientific methods to sample judges.
3. Select samples that are sufficient in size to provide stable average recommendations.
4. Use standard-setting procedures that are within the domain of knowledge and capabilities of sampled judges.
5. Train judges to understand and interpret information provided to them and to perform their tasks thoughtfully and consistently.
6. Inform judges about the requirements of the teaching jobs for which test performance will determine certification.
7. Provide judges with information on the relationship between test performance and job performance.

8. Provide judges with information on the distribution of test performances of a representative sample of examinees.
9. Provide judges with opportunities to reconsider their initial recommendations on test standards.

If these guidelines are followed, test standards will be produced by an appropriate and adequate sample of qualified, capable judges who understand their tasks and the implications of their recommendations. In setting test standards we can demand no more. And we should accept no less.

References

American Psychological Association, American Educational Research Association, National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement.* Washington, DC: American Council on Education.

Brennan, R. L. & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4,* 219-240.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnum, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cross, L. H., Impara, J. C., Frary, R. B. & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement, 21,* 113-129.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16,* 407-424.

Ebel, R. L. (1972). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor & Department of Justice (1978). Adoption by four agencies of uniform guidelines on employee selection procedures. *Federal Register, 43* (166), 38290-38315.

Glass, G. V (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237-261.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461-475.

Jaeger, R. M. (1986). Policy issues in standard setting for professional licensure tests. In W. P. Gorth & M. L Chernoff (Eds.), *Testing for teacher certification.* Hillsdale, NJ: Lawrence Erlbaum.

Jaeger, R. M. (in press). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement.* (3rd. ed.) New York: Macmillan.

Kane, M. & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. *Applied Psychological Measurement, 8,* 107-115.

Linn, R. L., Madaus, G. & Pedulla, J. (1982). Minimum competency testing: Cautions on the state of the art. *American Journal of Education,* November, 1-35.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Rebell, M. A. (1986). Disparate impact of teacher competency testing on minorities: Don't blame the test-takers -- or the tests. *Yale Law and Policy Review, IV,* 375-403.

Reilly, R. R., Zink, D. L. & Israelski, E. W. (1984). Comparison of direct and indirect methods for setting minimum passing scores. *Applied Psychological Measurement, 8,* 421-429.

Saunders, J. C., Ryan, J. P. & Huynh, H. (1981). A comparison of two approaches to setting passing scores based on the Nedelsky procedure. *Applied Psychological Measurement, 5,* 209-217.

Secolsky, C. (1987). Attitudinal and interpretive influences on difficulty ratings in setting standards for high school minimum competency tests. *Journal of Educational Research, 80,* 227-232.

Shepard, L. A. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement.* (pp. 59-71). Washington, DC: National Council on Measurement in Education.

Shepard, L. A. (1980a). Technical issues in minimum competency testing. In D. C. Berliner (Ed.), *Review of research in education, 8* (pp. 30-82). Washington, DC: American Educational Research Association.

Shepard, L. A. (1980b). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447-467.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction.* Baltimore, MD: Johns Hopkins University Press.

Van der Linden, W. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 4,* 295-305.

Zieky, M. J. & Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests.* Princeton, NJ: Educational Testing Service.